Perry, JL., Nicholls, AR., Clough, PJ. & Crust, L. 2015, Assessing Model Fit: Caveats and Recommendations for Confirmatory Factor Analysis and Exploratory Structural Equation Modeling. *Measurement in Physical Education and Exercise Science* vol 19, no. 1, pp. 12-21

Abstract

Despite the limitations of overgeneralizing cutoff values for confirmatory factor analysis (CFA; e.g., Marsh, Hau, & Wen, 2004), they are still often employed as golden rules for assessing factorial validity in sport and exercise psychology. The purpose of this study was to investigate the appropriateness of using the CFA approach with these cutoff values for typical multidimensional measures. Furthermore, we ought to examine how a model could be respecified to achieve acceptable fit and explored whether exploratory structural equation modeling (ESEM) provides a more appropriate assessment of model fit. Six multidimensional measures commonly used in sport and exercise psychology research were examined using CFA and ESEM. Despite demonstrating good validity in previous research, all eight failed to meet the cutoff values proposed by Hu and Bentler. ESEM improved model fit in all measures. In conclusion, we suggest that model misfit in this study demonstrates the problem with interpreting cutoff values rigidly. Furthermore, we recommend ESEM as a preferred approach to examining model fit in multidimensional measures.

Keywords

psychometrics, measurement, cutoff values, modification indices

Jöreskog (1969) developed confirmatory factor analysis (CFA) to examine psychometric models, and the use of CFA has risen exponentially in recent years and is particularly prominent in sport and exercise psychology. Searches on SPORTdiscus for the term “confirmatory factor analysis” in titles, keywords, and abstracts revealed that 180 papers employing CFA techniques were published from 1990 to 1999, compared to 549 papers from 2000 to 2009. The limitations of CFA approaches have been documented (e.g., Hopwood & Donnellan, 2010; Marsh, Hau, & Wen, 2004). However, there is still a lack of acknowledgment of these limitations in the sport and exercise psychology literature.

Theoretically, CFA represents an objective test of a theoretical model. In practice, conducting all factor analytic procedures requires a series of judgments. By far, the most important judgment made in CFA is whether a model is deemed to be acceptable or not. Logically, the process of accepting or rejecting models is fairly simple, in that the aim is to avoid concluding that a good model is bad, and that a bad model is good (MacCallum, Browne, & Sugawara, 1996). This is typically achieved by examining the absence or presence of misspecifications, which are errors between the prescribed model and the estimated parameters. In structural equation modeling, of which CFA is one form, the goodness of a model is typically determined by the absence (good) or presence (bad) of misspecifications (Saris, Satorra, & van der Veld, 2009). The clearest of all the parameters for making judgments on the acceptability of model fit is the chi-square ($\chi 2$). However, as initially observed by Bentler and Bonett (1980) and many thereafter (e.g., Saris et al., 2009), because this statistic is sensitive to sample size, it will reject models that have only a trivial misspecification, thus leading to increased type II error. The solution appears to be to use a selection-of-fit indices that calculate exact model fit based on chi-square (e.g., standardized root mean square residual or goodness-of-fit index), relative fit indices that compare the hypothesized model to an independent baseline model (e.g., Tucker-Lewis index or incremental fit index), and noncentrality-based indices that test the alternative hypothesis rather than the null (e.g., Bentler’s comparative fit index or the root mean square error of approximation).

Hu and Bentler (1999) proposed cutoff criteria for all commonly cited fit indices by examining rejection rates on hypothetical models. These proposed criteria, including Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) close to .95, Standardized Root Mean-square Residual (SRMR) of close to .08, and Root Mean Square Error of Approximation (RMSEA) of close to .06, are referred to as a matter of routine in studies using any kind of structural equation methods. While reference to Hu and Bentler's (1999) suggested cutoffs is not necessarily an issue itself, the extent to which many researchers view these recommendations as golden rules potentially creates an substantial amount of type II errors. In sport and exercise psychology, frequent judgments regarding the factorial validity of a measurement scale are made according to these rules. Marsh et al. (2004) keenly and accurately pointed out that Hu and Bentler offered caution about using such cutoff values and concisely explain the dangers of overgeneralizing the findings from Hu and Bentler in search of golden rules. Indeed, Marsh et al. referred to a traditional cutoff values amounting to "little more than rules of thumb based largely on intuition and have little theoretical justification" (2004, p. 321). Although scales are often published and used despite falling short of cutoffs, there are also examples of the psychometric properties of scales in sport and exercise psychology being dismissed as a result of adherence to cutoffs as golden rules. One popular example comes in the evolution of the sport motivation scale (SMS), originally developed by Pelletier et al. (1995), which was examined and revised by Mallett, Kawabata, Newcombe, Otero-Forero, and Jackson (2007). Despite the SMS originally being subjected to CFA, Mallett and colleagues conducted another CFA and found a model fit just short of the recommended criteria (CFI = .87, SRMR = .06, RMSEA = .06). The authors went on to describe the model fit as "poor" and used this as justification to claim that the scale required revision.

The use of CFA techniques for examining factorial validity and identifying acceptable levels of fit is certainly not straightforward. Hopwood and Donnellan (2010) illustrated the difficulty by examining eight common personality measurements. Hopwood and Donnellan applied more relaxed cutoff criteria than Hu and Bentler (1999; e.g., CFI and TLI > .90, RMSEA < .10) and allowed cross-loadings in some of the measures analyzed. Even so, by conducting CFAs, the authors found that none of the scales used came close to the recommended cutoff values. Interestingly, even the best-performing measure achieved a model fit well below the commonly accepted criteria, despite commonly being accepted as an appropriate assessment of personality. The length and complexity of personality measures means that employing the same requirements of such models compared to short, simple models is simply not appropriate. A CFA model typically constrains items to loading on only one factor as an independent cluster model (ICM) (CFA-ICM; Marsh et al., 2009), resulting in misspecification for each cross-loading. Long (i.e., many items), complex (i.e., many factors) measures therefore, have much less chance of achieving an acceptable fit. In providing their own caveat for using CFA, Hopwood and Donnellan (2010) described what they call The Henny Penny Problem after the character from the children's tale who lamented that the sky was falling after an acorn fell on his head. The authors pointed out that claims that a measure is invalid because of a weak CFA fit is exaggerated and ignores other types of validity such as content and criterion-related validity. Such personality assessments could perhaps perform better in a CFA by reducing their size and/or complexity, but if this is at cost of predictive or other forms of validity, it is simply not a virtuous academic pursuit.

When encountering misspecifications in a CFA model, the researcher has several options. They can either (a) determine that the misspecification is irrelevant and proceed, (b) concede that the

misspecification is significantly relevant and therefore reject the model, or (c) modify the model to achieve an acceptable fit. Such modification can be achieved using the modification indices (MI) provided in CFA output. The MI provide an estimate increase in the chi-square for each fixed parameter if it were to be freed. In ICMs, covariances between items from questionnaires are typically fixed to zero. By identifying significant MI and allowing them to be estimated, chi-square is reduced, thus yielded a better statistical model fit. The use of MI to respecify poorly fitting models was effectively demonstrated by MacCullum (1986) and further recommended by Saris, den Ronden, and Satorra (1987) and Saris et al. (2009). It should be noted however, that all of these authors also urge caution because this data driven approach does not necessarily hold any theoretical relevance. Indeed, MacCullum (1986) found that in half of the models tested in a simulation study, MI did not find a true model. Several authors (e.g., Kaplan, 2009) have referred to such respecification as atheoretical, claiming that it is merely capitalizing on chance within a sample. The process of using MI is seldom reported and therefore presumably seldom conducted in sport and exercise psychology.

Exploratory Structural Equation Modeling (ESEM) provides an alternative to CFA-ICM, which is effectively an integration of exploratory factor analysis (EFA) and CFA methods, which could be considered as a EFA-SEM approach (Asparouhov & Muthén, 2009). CFA-ICM assesses an a priori model that typically allows observed variables to load only onto their intended factor. Typically, all loadings, regardless of their significance, onto other latent variables are constrained to zero (Figure 1). This means that all trivial, non-significant cross-loadings will contribute to model misspecification (Ashton & Lee, 2007). This misspecification is defined by Hu and Bentler (1998, p. 427) as when "one or more parameters are fixed to zero where population values are non-zeros (i.e., an underparameterized misspecified model)." Clearly in many psychometric measures, particularly long, multidimensional scales, this can become a substantial issue. Moreover, questionnaires that are aggregated to enable an overall score to be derived as well as individual subscale scores to include appropriate internal consistency must have moderate to high inter-correlations, and therefore, many non-zero cross-loadings. A common example of such an aggregated measure is the Mental Toughness Questionnaire-48 (MTQ48; Clough, Earle, & Sewell, 2002). Church and Burke (1994) explained that ICMs are too restrictive for research where secondary or cross-loadings are likely, such as personality research. It is this reason why Hopwood and Donnellan (2010), and others before them, found such difficulty in obtaining a satisfactory CFA fit on personality scales. ESEM provides standard errors for all rotated parameters. As such, it allows all observed variables to load on all latent variables (Figure 2). This overcomes the issue of secondary, often non-significant cross-loadings causing irrelevant model misspecification, and therefore, the potential rejection of a good model. This was demonstrated by Marsh et al. (2010), who assessed the 60-item NEO Five-Factor Inventory using CFA and ESEM methods. The authors found that ESEM noticeably outperformed CFA in goodness of fit and construct validity.

FIGURE 1 An illustration of model structure with estimated parameters in confirmatory factor, independent clusters model analysis. Note: y represents the latent variables, which are typically subscales in self-report psychology measures, while x represents each observed variable, typically an item within a questionnaire, and e represents the residual error.

FIGURE 2 An illustration of model structure with estimated parameters in exploratory structural equation modeling.

Given the exponential rise in the use of CFA, it is crucial to examine the potential limitations of the technique, or the interpretation of CFA-ICM models using cutoff values. The purpose of this study was to firstly assess the likelihood that common quantitative measures in sport and exercise psychology can meet the cutoff values proposed by Hu and Bentler (1999) with independent samples. Secondly, we conducted ESEM on all scales to examine if this is likely to be a preferred alternative to CFA. We hypothesized that the majority of measurement scales used in the study would fall below the cutoff values proposed by Hu and Bentler (1999), and all chi-square values would suggest model misfit (i.e., $p < .001$). We also hypothesized that ESEM would provide a better model fit on all measurement scales, proportional to the amount of factors and whether or not the factors provide an aggregated score.

METHOD

Participants

We collated data from using six commonly used psychometric scales in sport and exercise psychology. The measures were selected to represent a range of complexities in terms of the number of items (22–48) and factors (3–10). The measures also represent a variety of interrelationships between subscales, where some have highly correlated subscales and others have relatively independent subscales. Participant information for each scale used is displayed in Table 1. All samples were gathered using athletes from a range of individual and team sports. For each sample, participants were recruited by approaching the head coach of a team, and all completed the questionnaire using pen and paper following informed consent. Where possible, heterogeneous samples were sought, as the measures examined in this article were largely validated on samples including both genders and from a range of backgrounds, sports, and performance levels.

TABLE 1 Demographic Details for Each Measurement Scale

Measures

Coping Inventory for Competitive Sport (CICS)

The CICS (Gaudreau & Blondin, 2002) examines 10 coping subscales using 39 items requiring a response on a five-point Likert-type scale anchored from 1 (Does not correspond at all to what I did or thought) to 5 (Corresponds very strongly to what I did or what I thought). For the purposes of this study, the CICS was only considered as a 10-factor model, and hierarchical models were not assessed. The CICS was developed using a sample of 316 French-Canadian athletes (54% male) aged 14 to 28 from a range of international (17%) to regional (35%) levels. Participants were drawn from a range of team and individual sports. Gaudreau and Blondin presented an acceptable CFA fit when the CICS was published (CFI = .93, TLI = .92, RMSEA = .04), also demonstrating sufficient concurrent

and divergent validity. Fletcher (2008) examined the psychometric properties of the CICS over a 10-week period, concluding that the measure is strong, obtaining meaningful and interpretable data.

Stress Appraisal Measure (SAM)

The SAM (Peacock & Wong, 1990) contains seven subscales with 28 items in total requiring a response on a five-point Likert-type scale anchored from 0 (Not at all) to 5 (Extremely). At the time of publication, Peacock and Wong presented support for the internal consistency and construct validity of the SAM. The SAM was developed in a series of studies using undergraduate students. While they did not conduct CFA, in developing and assessing the psychometric properties of a Turkish version, Durak and Senol-Durak (2013) presented a good model fit (CFI = .93, TLI = .92, SRMR = .06, RMSEA = .05). Durak and Senol-Durak tested the model on a sample of 461 undergraduate students (49.5% male) aged 17 to 33.

Mental Toughness Questionnaire-48 (MTQ48)

The MTQ48 (Clough et al., 2002) contains six subscales on 48 items requiring a response on a five-point Likert-type scale from 1 (Strongly disagree) to 5 (Strongly agree). Perry, Clough, Earle, Crust, and Nicholls (2013) found support for the factorial validity and reliability of the scale using a sample of over 8,000 from a variety of business, education, and sport backgrounds. The athlete sample (n = 442, 72.4% male) contained a range of sports and level of participation. The authors reported both CFA model fit (CFI = .85, TLI = .85, SRMR = .05, RMSEA = .05) and ESEM model fit (CFI = .94, TLI = .92, SRMR = .02, RMSEA = .03).

Sport Motivation Scale-6 (SMS-6)

The SMS-6 (Mallett et al., 2007) assesses a six-factor model of sport motivation on 24 items requiring a response on a seven-point Likert-type scale from 1 (Does not correspond at all) to 5 (Corresponds exactly). Mallett et al. examined the factorial validity of the scale on two large samples (n = 614 and 557, 44.2% male) from a range of individual and team sports. Mean age of performers was 20.0 and participation was at a variety of levels from international (19% and 7.1%) to recreational (43% and 16.5%). The authors claimed improved model fit (CFI = .93, TLI = .92, SRMR = .04, RMSEA = .05) compared to its earlier incarnation (The Sport Motivation Scale; Pelletier et al., 1995); the SMS-6 also demonstrated concurrent validity.

Sport Emotion Questionnaire (SEQ)

The SEQ (Jones, Lane, Bray, Uphill, & Catlin, 2005) examines five emotions using 22 items requiring a response on a five-point Likert-type scale from 0 (Not at all) to 5 (Extremely). Participants are asked to indicate the extent to which they experience each emotion at the time of completing the SEQ. At the time of publication, Jones et al. demonstrated reasonable model fit (CFI = .93, RMSEA = .07), concurrent and construct validity, and internal consistency. The sample used to examine factorial validity comprised of 518 athletes (57.9% male) with a mean age of 21.2 from a range of team and individual sports from varsity and regional competitions.

Coping Self-Efficacy Scale (CSES)

The CSES (Chesney, Neilands, Chambers, Taylor, & Folkman, 2006) consists of 26 items and three subscales requiring a response on an 11-point Likert-type scale from 0 (Cannot do at all) to 10 (Certain can do). In publishing the CSES, Chesney et al. presented satisfactory model fit (CFI = .95, SRMR = .05, RMSEA = .07), concurrent validity, and internal consistency. The development and testing of the CSES used non-sport samples of 348 men.

Procedure

All data were collected using pen and paper method in the presence of researchers to ensure authenticity. Ethical clearance was obtained by an ethics committee at a UK higher-education institution before data was collected. All participants were assured of confidentiality, were encouraged to complete questionnaires honestly, and gave informed consent.

Data Analysis

Preliminary analysis checked for missing data and outliers before univariate skewness and kurtosis and multivariate kurtosis were examined. CFA was conducted on all measurement scales using Mplus 7.0 (Muthén & Muthén, 2012). Model fit was assessed using chi square ($\chi 2$), the comparative fit index (CFI), the Tucker-Lewis index (TLI), standardized root mean square residual (SRMR), and root mean squared error of approximation (RMSEA). Chi-square and SRMR represented absolute fit indices, CFI and TLI provided incremental indices, and RMSEA presented a parsimony-adjusted measure. All analyses used the robust maximum likelihood method (MLR) with epsilon value .05, and the oblique geomin rotation, as factors in all models are theoretically correlated.

To examine how easily fixed a model could be, MI were used to correlate observed variables until a better model fit was found, using an iterative process as recommended by Oort (1998). In each analysis, the MI with the highest value were sequentially selected one at a time to enable observed error variables to correlate. Oort demonstrated that the process should be iterative, whereby only one modification is made at once, as others may contain biases based on the existing structure. This enabled firstly, to assess if this generated an acceptable model fit. Secondly, if it did, the amount of modifications required to achieve the fit were identified. However, this begins to deviate from the intended theoretical design of the original model. To assess if the model had deviated, the respecified model was cross-validated the respecified model by testing model fit on two random halves of the original sample. If there was a clear difference (ΔCFI > .1) between the model fits, the modified model was deemed to have failed cross-validation. Cross-validation using random half samples is a useful way of checking the extent to which model fit is sample-specific. This was used to determine the consistency of fit of CFA-ICM models. Furthermore, measurement invariance is a robust assessment of the stability of a model by assessing variance in factor loadings, intercepts, and factor means. This was used to determine if the measures used satisfied the assumption of invariance.

For all scales, ESEM was conducted, employing the same fit indices as CFA-ICM. For all ESEM analyses, the number of factors was identified, but items could load freely onto all factors. As ESEM could potentially achieve a good fit by finding unintended factor loadings, the model fit alone cannot be relied on without then examining the individual parameter estimates. To assess this, we computed the proportion of items that loaded significantly ($p < .05$) onto intended factors from the CFA-ICM, the number of significant cross-loadings, and the number of significant cross-loadings that

were greater than the loading onto the intended factor. It is important that intended factor loadings are substantially greater than latent factor loadings from unintended items, as subscale scores are derived from the sum of the intended only and do not consider unintended loadings. The Satorra-Bentler scaled chi-square difference test (Satorra, 2000) was used to identify if ESEM models produced statistically significantly different model fits to CFA-ICMs.

RESULTS

Preliminary Analyses

Less than 0.1% of data was missing in all samples, and there were no issues with outliers, following examination of Q-Q plots. All variables presented no issues with univariate skewness (< 2) or kurtosis (< 7). All multidimensional scales presented a departure from multivariate kurtosis. Consequently, the MLR was used in analyses.

Confirmatory Factor Analyses

A summary-of-fit indices from the CFAs are displayed in Table 2. It is worth noting that of the eight measurement scales assessed, all chi-square statistics results were statistically significant ($p < .001$). Moreover, none of the measures achieved cutoff values for CFI and TLI of > .95, as recommended by Hu and Bentler (1999). Indeed, the SEQ was the only measure to reach the sometimes applied more relaxed cutoff value of > .90 for CFI and TLI. While all met the recommended SRMR cutoff of < .08, only three of the eight achieved an RMSEA of < .06. With the exception of the CSES, all measures demonstrated a high proportion of items loading correctly onto their intended factor.

TABLE 2 Summary of Fit Indices for Measures Using CFA

Model respecification using MI significantly improved model it for each scale (Table 3). All chi-square values remained significant. To examine if modifications had deviated from the initial model, all samples were randomly split in half and tested using the respecified model. The results of this cross-validation are displayed in Table 4. For some measures, such as the CICS and SEQ, the modified model was successfully cross-validated, because no significant change in model fit was observed. For most of the measures, it appears that the use of the MI may deviate from the original model, though the extent to which this is theoretically substantial requires further investigation. It is worth noting that despite some changes in fit indices, all models achieved a reasonable model fit (i.e., CFI > .90) in both samples.

TABLE 3 Model Fits Using Modification Indices

TABLE 4 Model Fits Using Modification Indices for Cross-Validation

To determine if sampling effects may have been a cause of model misfit, we examined measurement invariance in all measures. To examine measurement invariance, the random half subsamples for each scale were subjected to a series of multigroup CFAs on increasingly constrained models. In four steps, we assessed configural invariance (items of a scale are indicators of the same factors in different groups), metric invariance (factor loadings are equal across groups), scalar invariance (loadings and intercepts of the items that form a latent construct are invariant across groups), and factor means invariance (loadings, intercepts, and means are invariant across groups). To examine

significant changes across groups, we employed Cheung and Rensvold's (2002) recommendation of observing ΔCFI < .01. The results of these analyses are presented in Table 5. With the exception of the CD-RISC, all measures demonstrated strong measurement invariance.

TABLE 5 Fit Indices for Multigroup CFA

Exploratory Structural Equation Modeling

All multidimensional measurement scales presented significantly improved model fit using ESEM (see Table 6). Chi-square difference testing using the Satorra-Bentler scaled chi-square found all improvements in model fit between CFA-ICM and ESEM to be statistically significant ($p < .001$). On average, CFI increased by .082, TLI increased by .070, SRMR reduced by .032, and RMSEA reduced by .018. All chi-square significance values remained significant ($p < .001$). Despite the marked improvements in model fit, only four of the six scales presented a CFI ≥ .95, and none presented a TLI ≥ .95. All SRMR were < .08, and all RMSEA were below < .06. This indicates an inconsistency in Hu and Bentler's (1999) proposed cutoff values, as some appear to be much more achievable than others.

TABLE 6 Summary of Fit Indices for Measures Using ESEM

ESEM loadings were examined to assess whether they have loaded onto what would be their intended factor in a CFA-ICM. Furthermore, statistically significant cross-loadings or cross-loadings greater than the loading onto the intended factor represent a misspecification in the model. Approximately 90% of items loading onto their intended factor appears to be the norm, allowing for some cross-loadings. As expected, the only aggregated measure, the MTQ48, included a greater number of significant cross-loadings. Consequently, the increase in model fit for this measure between CFA and ESEM was greater.

DISCUSSION

The purpose of this study was, firstly, to assess the likelihood that common quantitative measures in sport and exercise psychology can meet proposed cutoff values and, secondly, to evaluate the ability of ESEM to provide a more appropriate estimate of model fit than CFA. Overall, the results indicate that a host of commonly-used scales fail to meet the cutoff values proposed by Hu and Bentler (1999). Respecification of the measurement models significantly improves model fit, and this is an option for researchers encountering misspecifications. Measurement invariance testing supports the stability of the scales used.

The results suggest that Hu and Bentler's (1999) proposed, and often implemented, cutoff values for a host of fit indices are unrealistic for many complex measures to achieve on a sample independent from that with which they were developed. While values for SRMR and RMSEA were achievable in standard CFA-ICM, modified CFA, and ESEM models, CFI reached the cutoff of .95 less frequently, and TLI seldom reached values close to .95. Furthermore, all analyses in this article yielded a significant chi-square, highlighting the hypersensitivity of chi-square as an assessment model of fit. When considered with previous demonstrations of validity, it appears that the primary reasons for model misfit are that (a) some cutoff values are overly strict for the measurement models, or (b) CFA-ICM is an inappropriate technique for analyzing the factorial validity of complex scales. Consequently, we urge caution for researchers when employing the CFA-ICM technique. Of course,

there are many examples of multidimensional scales achieving a good model fit on an independent sample, but if a scale fails to achieve such a fit, it should not be automatically discredited. As a minimum, researchers should acknowledge the limitations of the approach and rigid cutoff values to prevent the "Henny Penny" problem described by Hopwood and Donnellan (2010). Those referring to Hu and Bentler's (1999) suggested cutoff values as golden rules when conducting CFA on complex, multidimensional models would be well advised to review the hypothetical models used in the original paper to establish such cutoffs. Hu and Bentler (1999) presented a simple model that contained 15 observed variables and three factors. Each factor had five loadings of .70 to .80, and all cross-loadings were fixed to zero. Furthermore, they examined a "complex" model that enabled just three cross-loadings across the same matrix. This is a long way from the complexity of many of the measures commonly used in sport and exercise psychology and another example of the dangers in overgeneralizing Hu and Bentler's (1999) findings, a topic discussed in much greater depth by Marsh et al. (2004). Evidence of the issue with model complexity and Hu and Bentler's (1999) proposed cutoff values can be found by examining the variety of performance of scales against different fit indices. The RMSEA and TLI contain a penalty for model complexity using the ratio between chi-square and degrees of freedom. SRMR has no penalty for model complexity and was therefore able to meet the recommend cutoffs in all models.

The extent to which a misspecified model can be fixed remains contentious. From purely a statistical point of view, it is feasible to respecify the model using the MI. However, caution is urged when conducting this method, as all respecifications must be theoretically acceptable. This could be an acceptable approach as long as restrictions are placed on permissible modifications (MacCullum, 1986). Said differently, researchers should determine whether it is theoretically plausible for model respecification. An example might be freeing parameters between items within the same subscale or perhaps creating a higher-order model that allows covariances between some subscale items that are theoretically related. Researchers should also be aware of the potential for MI to aid the theoretical development of a measurement model.

ESEM is an emerging technique that is used either supplementary with CFA-ICM or instead of CFA-ICM. There are several studies that utilize ESEM effectively for the development and/or validation of a multidimensional measure outside of the sport domain (e.g., Marsh et al., 2010; Marsh, Nagengast, Morin, Parada, Craven, & Hamilton, 2011). In this study, we have demonstrated that this technique is a desirable alternative to CFA using scales frequently used in a sport context. Other than rare exceptions (e.g., Morin & Maïno, 2011), the use of ESEM in the sport and exercise psychology literature is limited at present. It was interesting to note that the SMS-6—which Mallett et al. (2007) developed to replace the original SMS (Pelletier et al., 1995) after suggesting that a model fit of CFI = .87, SRMR = .06, RMSEA = .06 was "poor"—presented a very similar model fit under the CFA-ICM analysis (CFI = .88, SRMR = .06, RMSEA = .07). However, under the ESEM analysis, the scale performed much better (CFI = .96, SRMR = .02, RMSEA = .05). Strict adherence to the common cutoff recommendations would mean we would have dismissed the SMS-6 model fit as poor, as Mallett et al. (2007) did with the original SMS. In reality, the SMS-6 has demonstrated consistently good factorial validity.

We propose that researchers should make a theoretical judgment on the appropriateness of the technique. For true ICMs in which subscales within the model are theoretically unrelated to each other or even opposed to each other, CFA should provide an accurate representation of the model

fit. If encountering misspecifications, researchers may consider the use of MI to improve model fit as long as they are able to theoretically justify their respecifications. The vast majority of multidimensional scales in sport and exercise psychology, however, are not true ICMs, because we can logically expect to find secondary loadings, particularly within highly correlated subscales or aggregated subscales. Under these circumstances, ESEM provides a more appropriate assessment of model fit than CFA and should be used from the outset, either in place of, or supplementary to, CFA. It is important to note, however, that ESEM should not be seen as a simple statistical fix to achieve higher CFI. By estimating more parameters, model fit inevitably improves. Regardless of the model fit indices, researchers must examine factor loadings to determine if the goodness of fit is derived from theoretical sound factor loadings, and not merely through the estimation of many parameters.

The variety of measures and relatively large sample sizes used is a strength of this article. There are however, some limitations to acknowledge. Firstly, the extent to which MI substantially change each model requires further investigation, as cross-validation was provided only by splitting the original sample. A true measure of this would be to improve a model fit using the MI on one large sample and then use a completely independent sample to cross validate the new model. Secondly, although we have selected a varying range of measures regarding length, factors, and aggregated scores, this is merely a sample of the many that are routinely published and used in research. All of the measures used were validated on samples similar to those used in this article with the exception of the CSES (Chesney et al., 2006). The CSES was developed and validated on a sample of males drawn from outside of sport. Consequently, potential misspecifications could result from demographic differences for this scale. Finally, it is worth noting that ESEM in this article has been solely used to estimate measurement models. The extent to which ESEM provides a solution to limitations in structural models requires further examination, perhaps in contrast to the use of parceling techniques.

In summary, we have demonstrated here that the proposed cutoff values by Hu and Bentler (1999) are unrealistic for most commonly used scales in sport and exercise psychology. That none of the measures used achieved the suggested cutoff values leads us to one of three conclusions: All of the measures we assessed are inadequate, CFA-ICM is not the most appropriate technique to examine their factorial validity, or the cutoff values are not appropriate. Because all of the measures used have previously provided evidence of their suitability, to accept the former is likely to lead to the rejection of many highly useful self-report measures. We feel the latter two conclusions provide a more true, progressive, and helpful way forward. Furthermore, we recommend that researchers examining more complex, multidimensional, or aggregated models should conduct ESEM in place of, or supplementary to, CFA.

Notes

Note: CICS = coping inventory for competitive sport; SAM = stress appraisal measure; MTQ48 = mental toughness questionnaire-48; SMS-6 = sport motivation scale-6; SEQ = sport emotion questionnaire; CSES = coping self-efficacy scale.

Note: CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean square residual; RMSEA = root mean square error of approximation; CICS = coping inventory for competitive sport; SAM = stress appraisal measure; MTQ48 = mental toughness questionnaire-48; SMS-6 = sport motivation scale-6; SEQ = sport emotion questionnaire; CSES = coping self-efficacy scale.

Note: CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean square residual; RMSEA = root mean square error of approximation; CICS = coping inventory for competitive sport; SAM = stress appraisal measure; MTQ48 = mental toughness questionnaire-48; SMS-6 = sport motivation scale-6; SEQ = sport emotion questionnaire; CSES = coping self-efficacy scale.

Note: CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean square residual; RMSEA = root mean square error of approximation; CICS = coping inventory for competitive sport; SAM = stress appraisal measure; MTQ48 = mental toughness questionnaire-48; SMS-6 = sport motivation scale-6; SEQ = sport emotion questionnaire; CSES = coping self-efficacy scale.

Note: All $\chi 2$ $p < .001$.

Note: CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean square residual; RMSEA = root mean square error of approximation; CICS = coping inventory for competitive sport; SAM = stress appraisal measure; MTQ48 = mental toughness questionnaire-48; SMS-6 = sport motivation scale-6; SEQ = sport emotion questionnaire; CSES = coping self-efficacy scale; Cross loadings were deemed statistically significant if $p < .05$.

REFERENCES

1. Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. Personality and Social Psychology Review, 11, 150–166. doi:10.1177/1088868306294907

2. Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. Structural Equation Modeling, 16, 397–438. doi:10.1080/10705510903008204

3. Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88, 588–606. doi:10.1037/0033-2909.88.3.588

4. Chesney, M. A., Neilands, T. B., Chambers, D. B., Taylor, J. M., & Folkman, S. (2006). A validity and reliability study of the Coping Self-Efficacy Scale. British Journal of Health Psychology, 11, 421–437. doi:10.1348/135910705X53155

5. Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of fit indexes for testing measurement invariance. Structural Equation Modeling, 9, 233–255. doi:10.1207/S15328007SEM0902_5

6. Church, A. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the big 5 and Tellegens 3-dimensional and 4-dimensional models. Journal of Personality and Social Psychology, 66, 93–114. doi:10.1037/0022-3514.66.1.93

7. Clough, P., Earle, K., & Sewell, D. (2002). Mental toughness: The concept and its measurement. In I. Cockerill (Ed.), Solutions in sport psychology, (pp. 32–43). London, UK: Thomson.

8. Durak, M., & Senol-Durak, E. (2013). The development and psychometric properties of the Turkish version of the Stress Appraisal Measure. European Journal of Psychological Assessment, 29, 64–71. doi:10.1027/1015-5759/a000079

9. Fletcher, R. (2008). Longitudinal factorial invariance, differential, and latent mean stability of the Coping Inventory for Competitive Sport. In M. P. Simmons & L. A. Foster (Eds.) Sport and exercise psychology research advances, (pp. 293–306). New York, NY: Nova Science Publishers Inc.

10. Gaudreau, P., & Blondin, J-P. (2002). Development of a questionnaire for the assessment of coping strategies employed by athletes in competitive sport settings. Psychology of Sport and Exercise, 3, 1–34. doi:10.1016/S1469-0292(01)00017-6

11. Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? Personality and Social Psychology Review, 14, 332–346. doi:10.1177/1088868310361240

12. Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modelling: Sensitivity to underparameterized model misspecification. Psychological Methods, 3, 424–453.

13. Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1–55. doi:10.1080/10705519909540118

14. Jones, M. V., Lane, A. M., Bray, S. R., Uphill, M., & Catlin, J. (2005). Development and validation of the Sport Emotion Questionnaire. Journal of Sport and Exercise Psychology, 27, 407–431.

15. Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34, 183–202. doi:10.1007/BF02289343

16. Kaplan, D. (2009). Structural equation modeling: Foundations and extensions (2nd ed.). Thousand Oaks, CA: Sage. OpenURL University of Lincoln

17. MacCullum, R. (1986). Specification searches in covariance structure modeling. Psychological Bulletin, 100, 107–120.

18. MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modelling. Psychological Methods, 1, 130–149. doi:10.1037//1082-989X.1.2.130

19. Mallett, C. J., Kawabata, M., Newcombe, P., Otero-Forero, A., & Jackson, S. (2007). Sport Motivation Scale-6 (SMS-6): A revised six-factor sport motivation scale. Psychology of Sport and Exercise, 8, 600–614. doi:10.1016/j.psychsport.2006.12.005

20. Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu & Bentler's (1999) findings. Structural Equation Modeling, 11, 320–341. doi:10.1207/s15328007sem1103_2

21. Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big-five factor structure through exploratory structural equation modeling. Psychological Assessment, 22, 471–491. doi:10.1037/a0019227

22. Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. Structural Equation Modeling, 16, 439–476. doi:10.1080/10705510903008220 [Taylor & Francis Online], [Web of Science ®]OpenURL University of Lincoln

23. Marsh, H. W., Nagengast, B., Morin, A. J. S., Parada, R. H., Craven, R. G., & Hamilton, L. R. (2011). Construct validity of the multidimensional structure of bullying and victimization: An application of exploratory structural equation modeling. Journal of Educational Psychology, 103, 701–732. doi:10.1037/a0024122 [CrossRef], [Web of Science ®]OpenURL University of Lincoln

24. Morin, A. J. S., & Maïno, C. (2011). Cross-validation of the short form of the physical self-inventory (PSI-S) using exploratory structural equation modeling (ESEM). Psychology of Sport and Exercise, 12, 540–554. doi:10.1016/j.psychsport.2011.04.003 [CrossRef], [Web of Science ®]OpenURL University of Lincoln

25. Muthén, L. K., & Muthén, B. O. (1998–2012). Mplus user's guide (7th ed.). Los Angeles, CA: Muthén & Muthén. OpenURL University of Lincoln

26. Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. Structural Equation Modeling: A Multidisciplinary Journal, 5, 107–124. doi:10.1080/10705519809540095 [Taylor & Francis Online], [Web of Science ®]OpenURL University of Lincoln

27. Peacock, E., & Wong, P. (1990). The stress appraisal measure (SAM): A multidimensional approach to cognitive appraisal. Stress Medicine, 6, 227–236. doi:10.1002/smi.2460060308 [CrossRef], [CSA]OpenURL University of Lincoln

28. Pelletier, L. G., Fortier, M. S., Vallerand, R. J., Tuson, K. M., Brière, N. M., & Blais, M. R. (1995). Toward a new measure of intrinsic motivation, extrinsic motivation, and amotivation in sports: The Sport Motivation Scale (SMS). Journal of Sport and Exercise Psychology, 17, 35–54. [Web of Science ®], [CSA]OpenURL University of Lincoln

29. Perry, J. L., Clough, P. J., Earle, K., Crust, L., & Nicholls. A. R. (2013). Factorial validity of the Mental Toughness Questionnaire 48. Personality and Individual Differences, 4, 587–592. doi:10.1016/j.paid.2012.11.020 [CrossRef], [Web of Science ®]OpenURL University of Lincoln

30. Saris, W. E., den Ronden, J., & Satorra, A. (1987). Testing structural equation models. In P. Cuttance & R. Ecob (Eds.), Structural modeling by example (pp. 202–220). New York, NY: Cambridge University Press. OpenURL University of Lincoln

31. Saris, W. E., Satorra, A., & van der Veld, W. (2009). Testing structural equation models or detection of misspecifications? Structural Equation Modeling, 16, 561–582. doi:10.1080%2F10705510903203433 [Taylor & Francis Online], [Web of Science ®]OpenURL University of Lincoln

32. Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra, (Eds.), Innovations in multivariate

statistical analysis. A Festschrift for Heinz Neudecker (pp. 233–247). London, UK: Kluwer Academic Publishers.